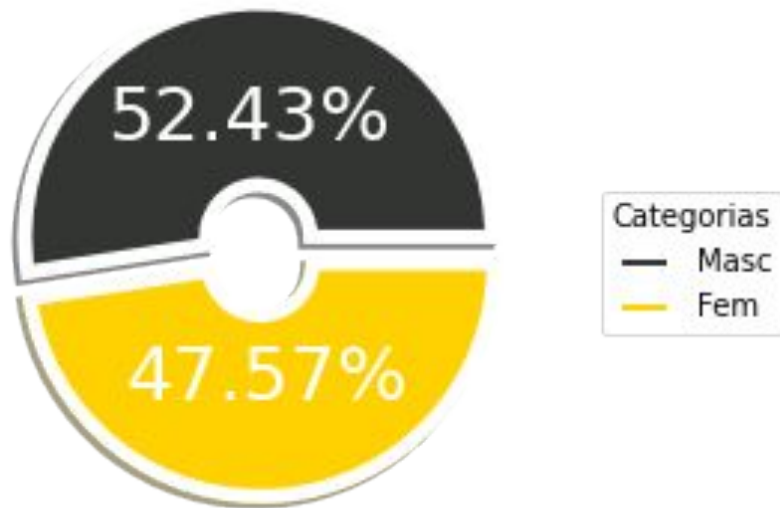


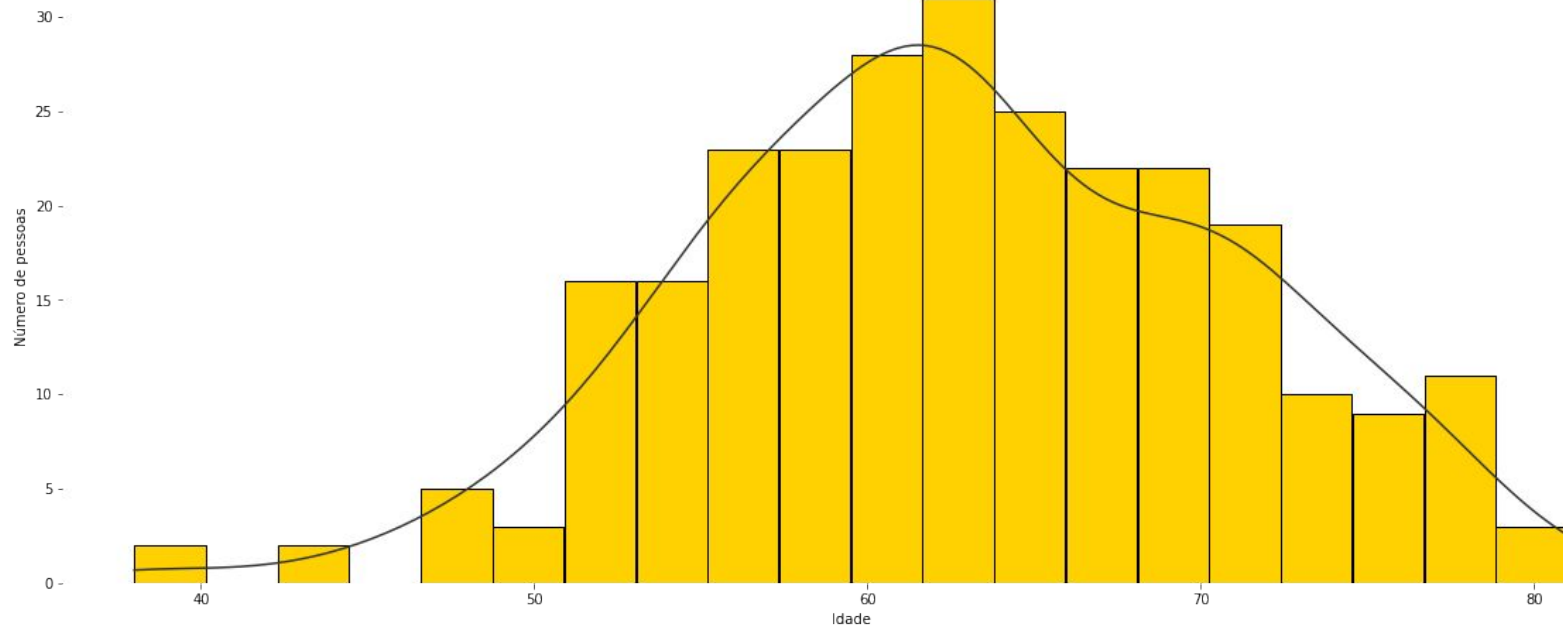
Como os dados estão distribuídos?

Distribuição entre os gêneros



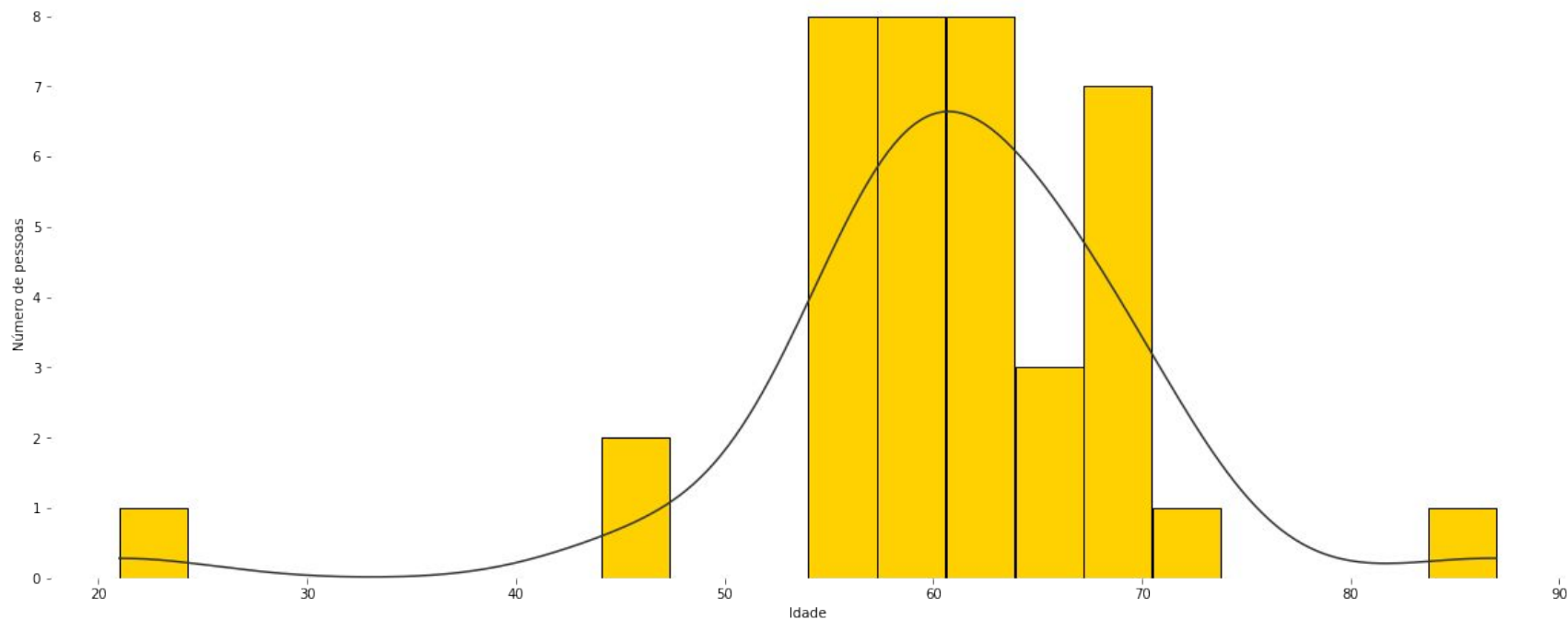
Como os dados estão distribuídos?

Distribuição dos casos positivos

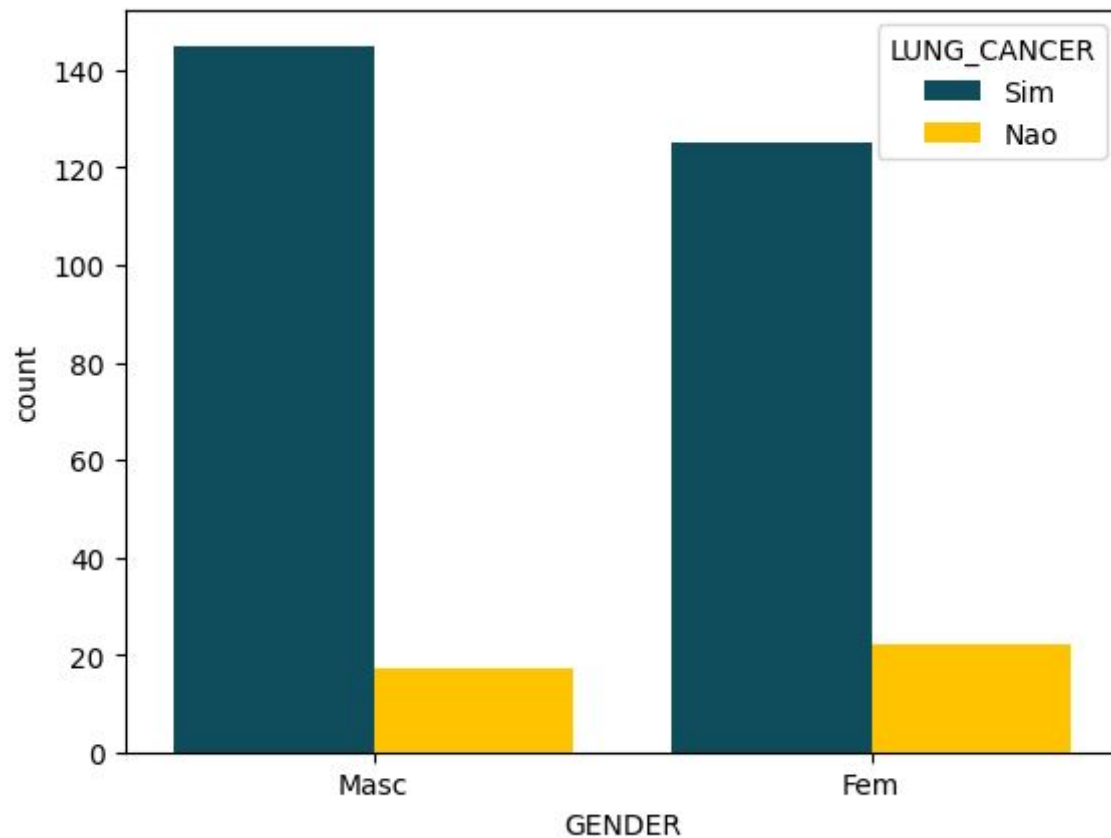


Como os dados estão distribuídos?

Distribuição dos casos negativos



Como os dados estão distribuídos?

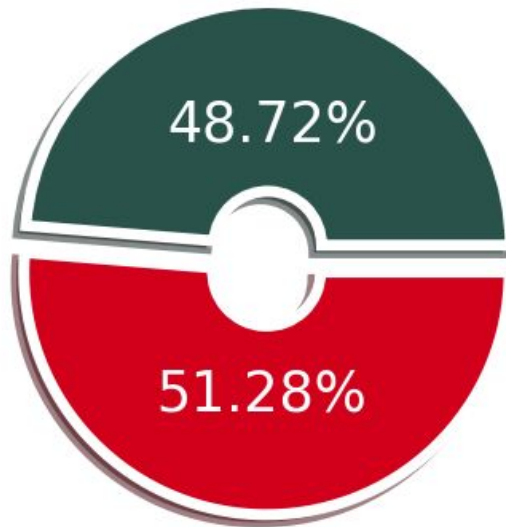


Como os dados estão distribuídos?

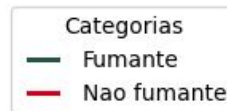
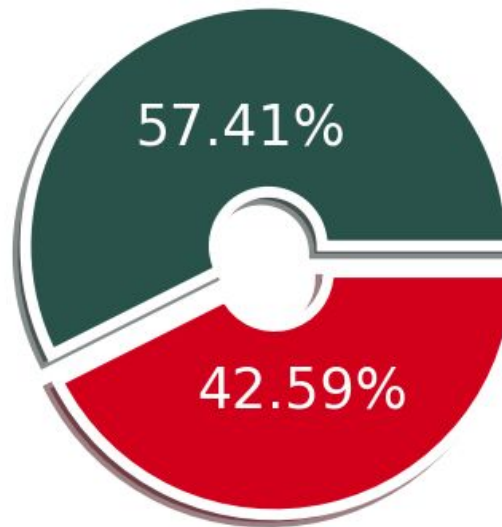
- Distribuição equilibrada para classe 'GENDER';
- Distribuição equilibrada para casos positivos e negativos entre 'GENDER';
- Desbalanceado para casos positivos e negativos;
- Idade centralizada em torno de 60 anos.

Quais dados são relevantes para indicação?(nesse dataset)

Fumantes para casos negativos

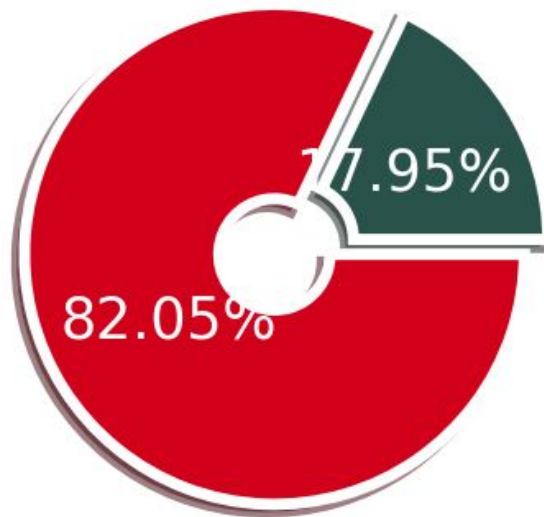


Fumantes para casos positivos

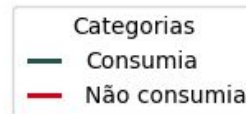
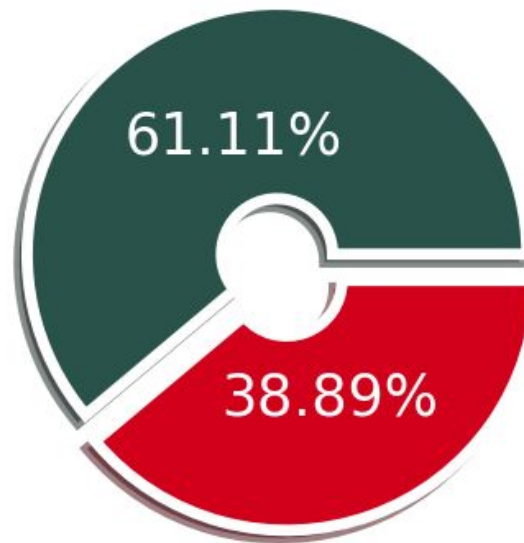


Quais dados são relevantes para indicação?(nesse dataset)

Consumo de alcool para casos negativos



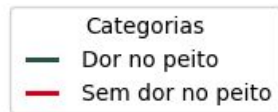
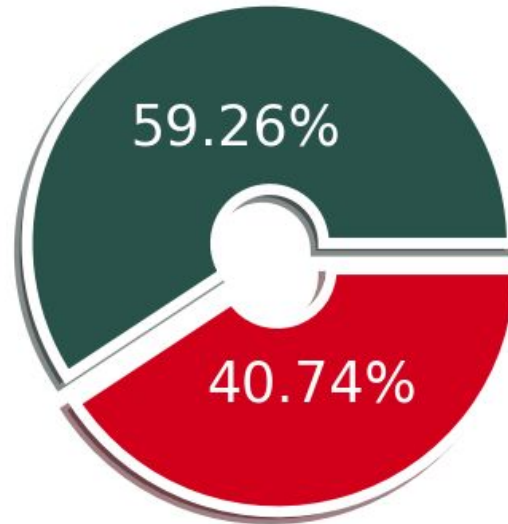
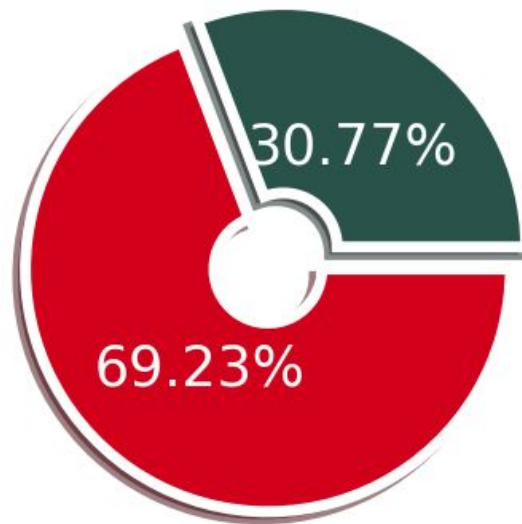
Consumo de alcool para casos positivos



Quais dados são relevantes para indicação?(nesse dataset)

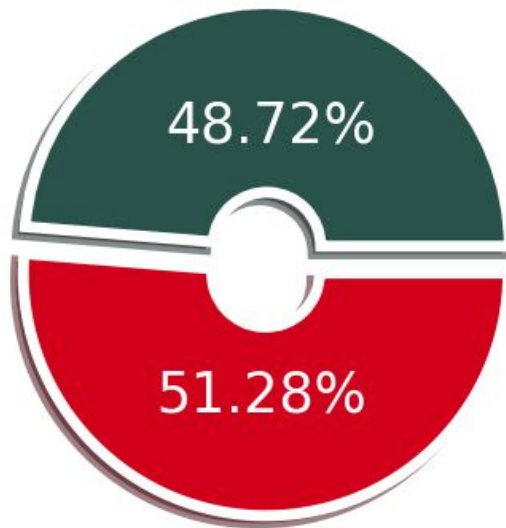
Dor no peito para casos negativos

Dor no peito para casos positivos

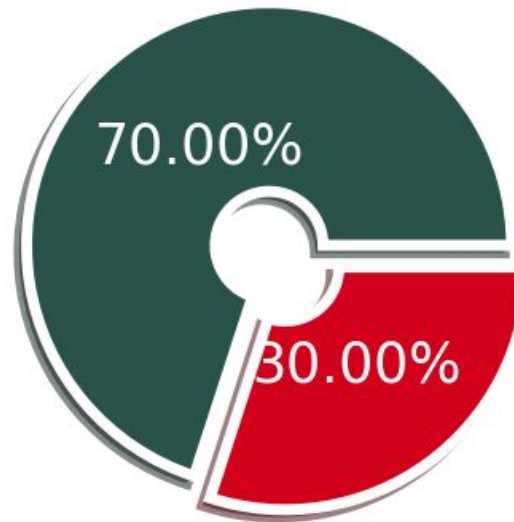


Quais dados são relevantes para indicação?(nesse dataset)

Fatiga para casos negativos



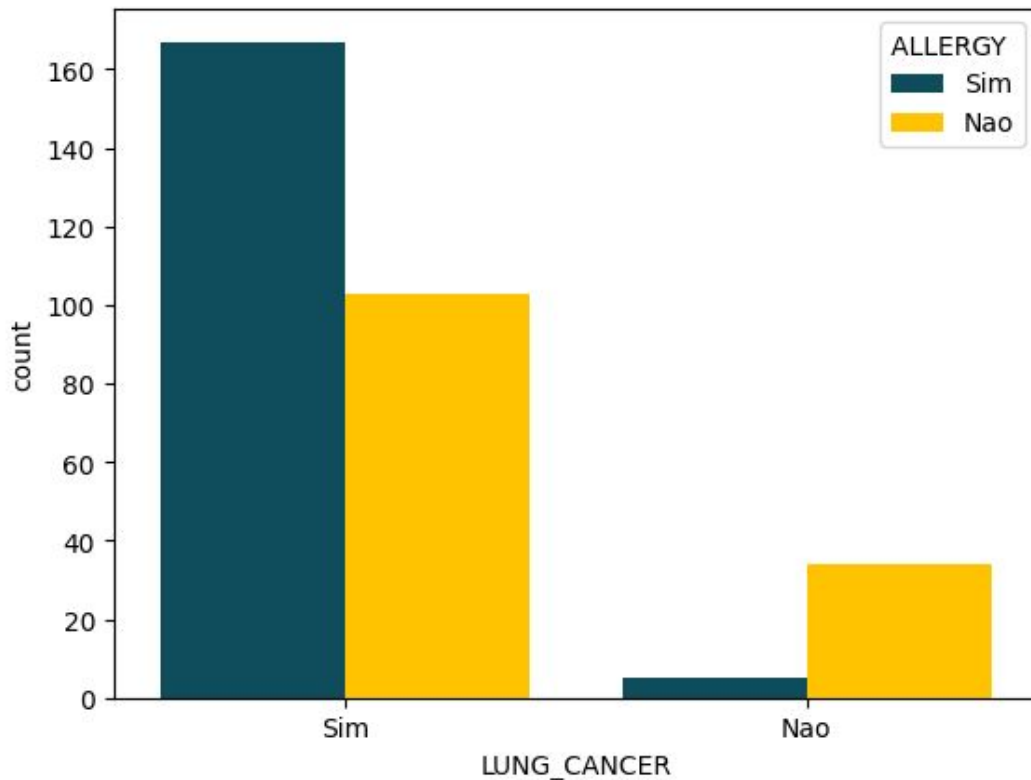
Fatiga para casos positivos



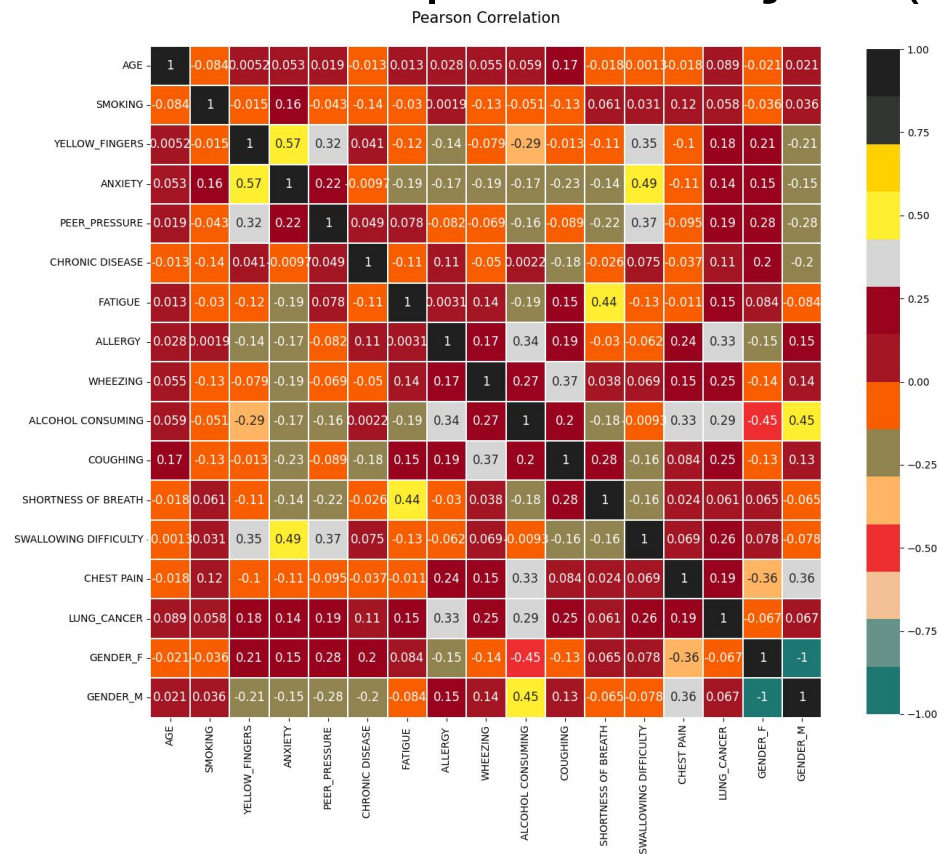
Quais dados são relevantes para indicação?(nesse dataset)

- Devido a natureza do dataset, a característica de fumante não mostrou ser um fator indicador para câncer no pulmão, obviamente isso não pode ser generalizado;
- Fatores como fadiga e alcoolismo se mostraram bem mais importantes.

Quais dados são relevantes para indicação?(nesse dataset)

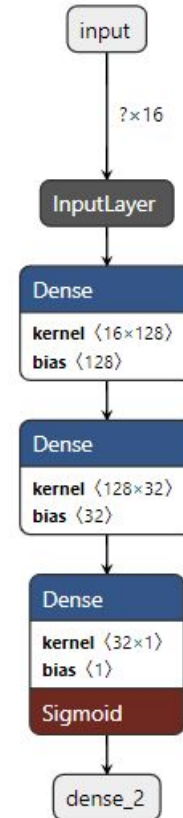


Quais dados são relevantes para indicação?(nesse dataset)



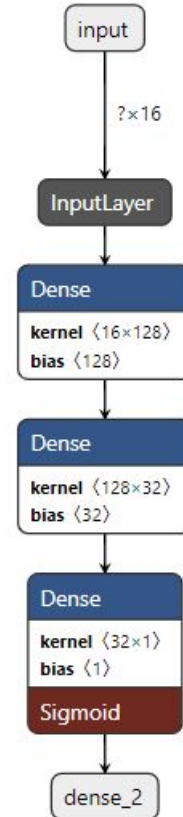
A remoção de características menos correlatadas afeta a classificação?

```
Shape of training data : (247, 16), (247,)
Shape of testing data  : (62, 16), (62,)
```



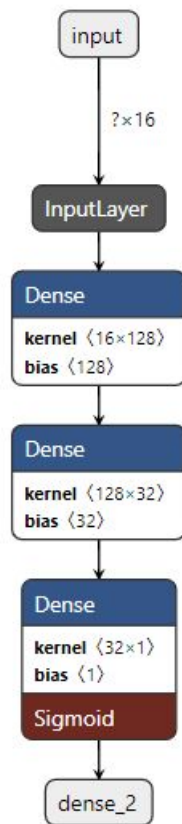
A remoção de características menos correlatadas afeta a classificação?

| | FATIGUE | ALLERGY | ALCOHOL CONSUMING | CHEST PAIN | LUNG_CANCER |
|---|---------|---------|-------------------|------------|-------------|
| 0 | 2 | 1 | 2 | 2 | 1 |
| 1 | 2 | 2 | 1 | 2 | 1 |
| 2 | 2 | 1 | 1 | 2 | 0 |
| 3 | 1 | 1 | 2 | 2 | 0 |
| 4 | 1 | 1 | 1 | 1 | 0 |

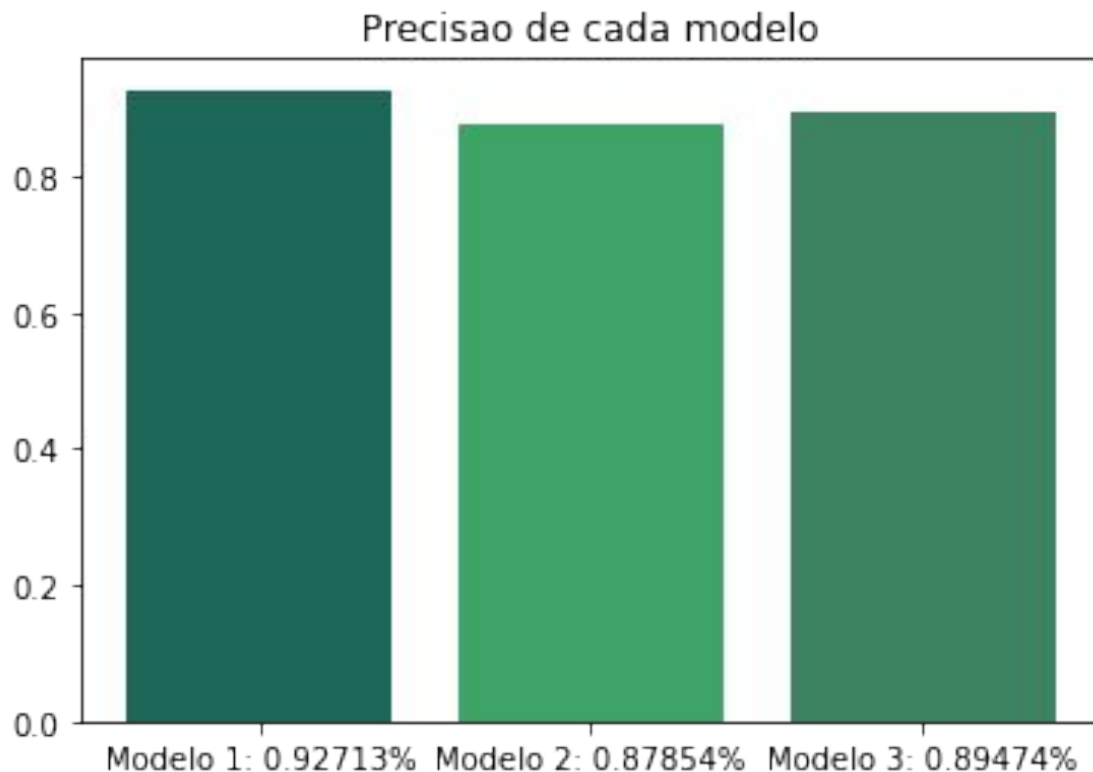


A remoção de características menos correlatadas afeta a classificação?

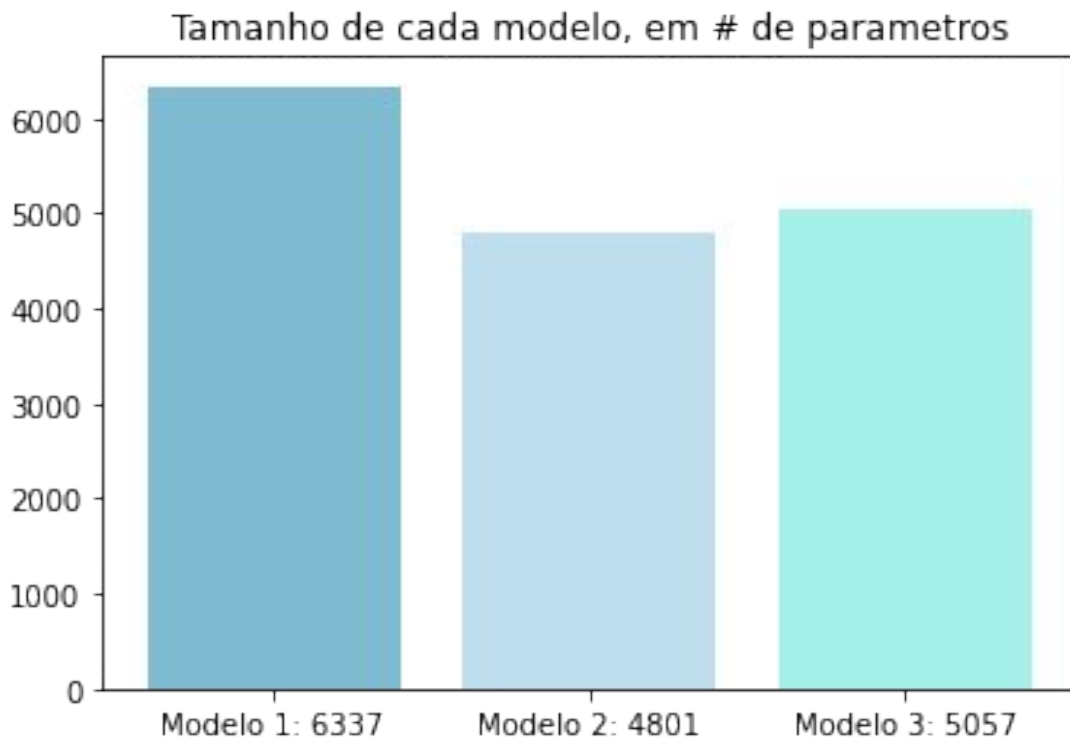
| | FATIGUE | ALLERGY | ALCOHOL CONSUMING | COUGHING | SWALLOWING DIFFICULTY | CHEST PAIN | LUNG_CANCER |
|---|---------|---------|-------------------|----------|-----------------------|------------|-------------|
| 0 | 2 | 1 | 2 | 2 | 2 | 2 | 1 |
| 1 | 2 | 2 | 1 | 1 | 2 | 2 | 1 |
| 2 | 2 | 1 | 1 | 2 | 1 | 2 | 0 |
| 3 | 1 | 1 | 2 | 1 | 2 | 2 | 0 |
| 4 | 1 | 1 | 1 | 2 | 1 | 1 | 0 |



A remoção de características menos correlatadas afeta a classificação?



A remoção de características menos correlatadas afeta a classificação?



A remoção de características menos correlatadas afeta a classificação?

- Conseguimos perceber que, com um custo de reduzir a precisão em 4% (~93% -> 89%) conseguimos uma redução de 20,2% no tamanho do modelo, em dados brutos isso pode ser visto melhor em grandes modelos como DALL-E 2 com 6 Bilhões de parâmetros.